Using Meta-Layers for GIS Dataset Tracking and Management

NEARC May 11, 2010

Brian Hebert

May, 2010

bhebert@scribekey.com

Presentation Outline

- Geospatial Metadata Today
- Issues and Challenges
- Using a RDB Metadata Repository (MR)
- Building and Loading the MR
- The Geospatial Difference
- Q&A

Project Background/Requirements

- Generate FGDC Metadata for Large Commercial Data Provider(s)
- Helping Federal Agency Collect, Describe, and Deliver Large Datasets
- Applies to any large collection of GIS layers, public or private.

200+ Countries
72 Layers
100s of Attributes
100s of Domains
Quarterly Updates

50+ States
400 Layers
1000s of Attributes
100s of Domains
Annual Updates

Typical Metadata Workflow

- Understand standard(s), XML, UML
- Find examples
- Use metadata editor
- Package metadata with data
- End user views metadata with ArcCatalog
- Metadata loaded onto server to support search/download

Geospatial Metadata Standards

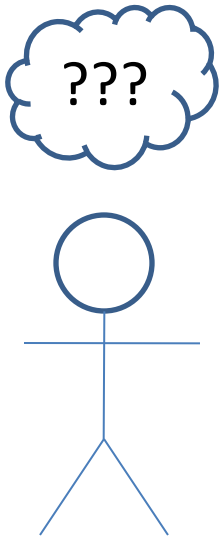
- FGDC CSDGM
  1990

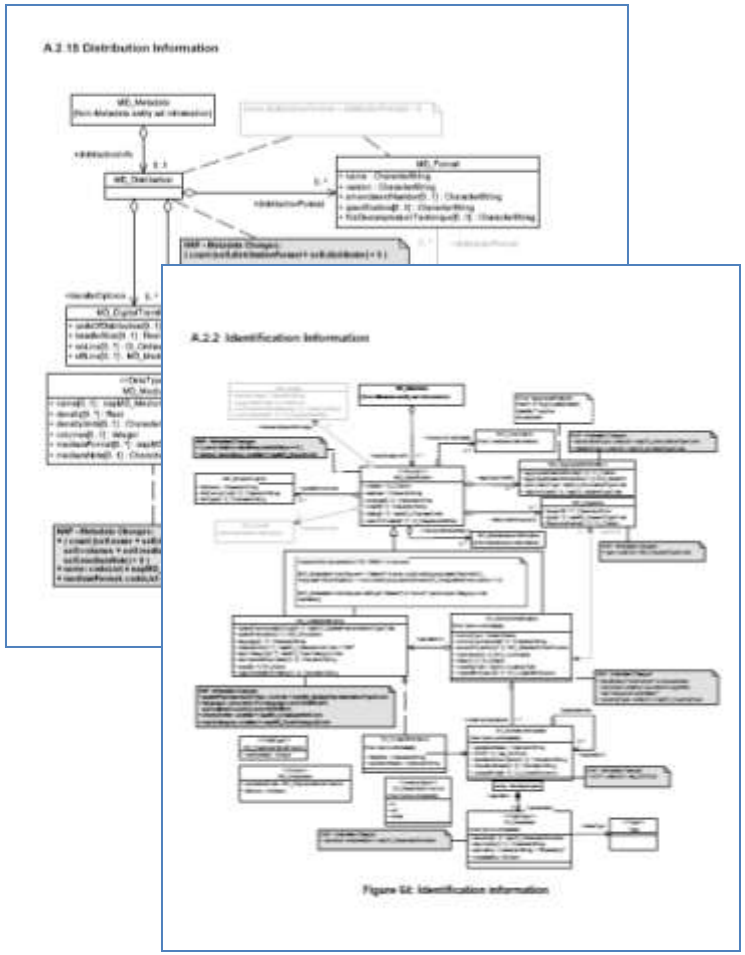
- ISO 19115/19110
  NAP 2009

- Both persisted as
  XML/XSD,
  rendered as HTML

Understanding the Standards

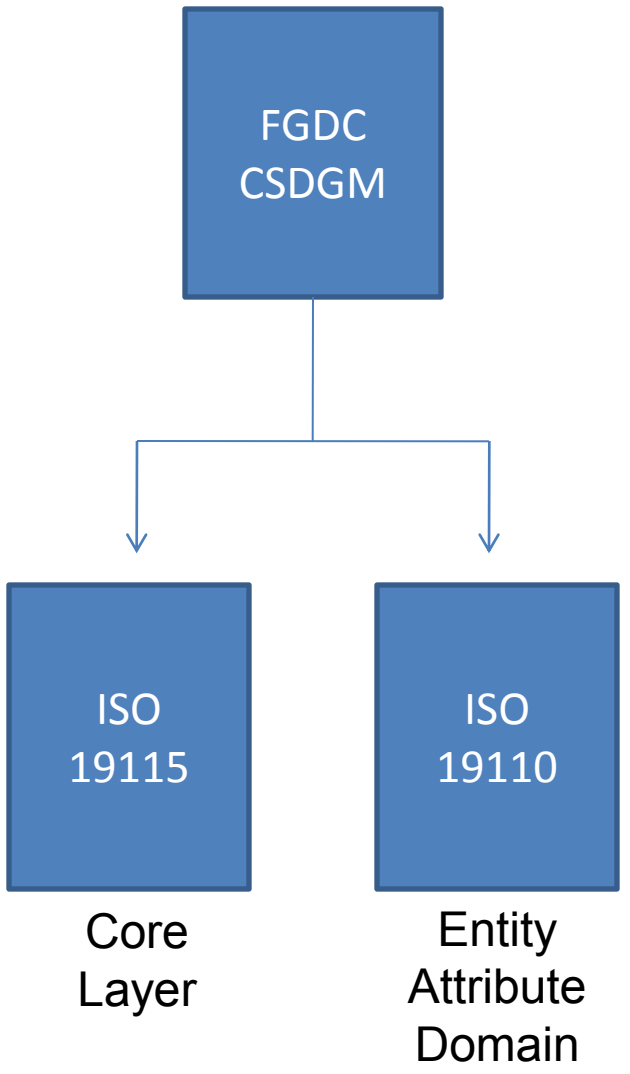
- Complicated for end users, what's core?
- UML, XSD, Grammar Production Rules
- Language and artifacts familiar to professional data modelers, academicians
- **No separation between logical and physical** as typically used in IT solution development





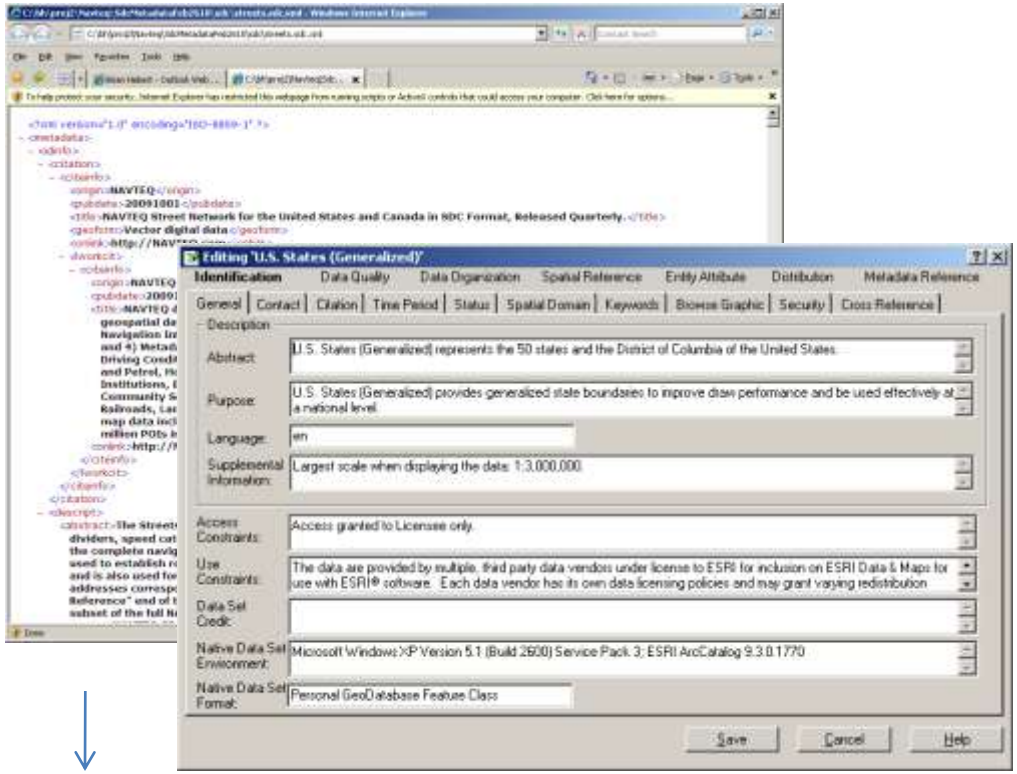
ISO 19115/19110 Split

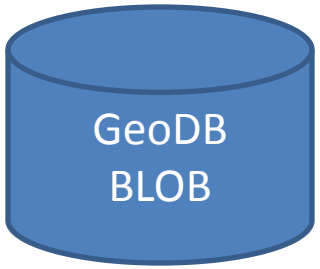
- FGDC CSDGM included both core metadata about layer and entity, attribute, domain info
- ***ISO 19115 Geographic Information – Metadata*** doesn't include this basic database-centric metadata
- ***19110 Feature Catalog*** does contain entity, attribute, domain info, but mixed in with a great deal of other material
- Presents a significant challenge to migration, for tool providers and users alike



Provider: Getting the Job Done

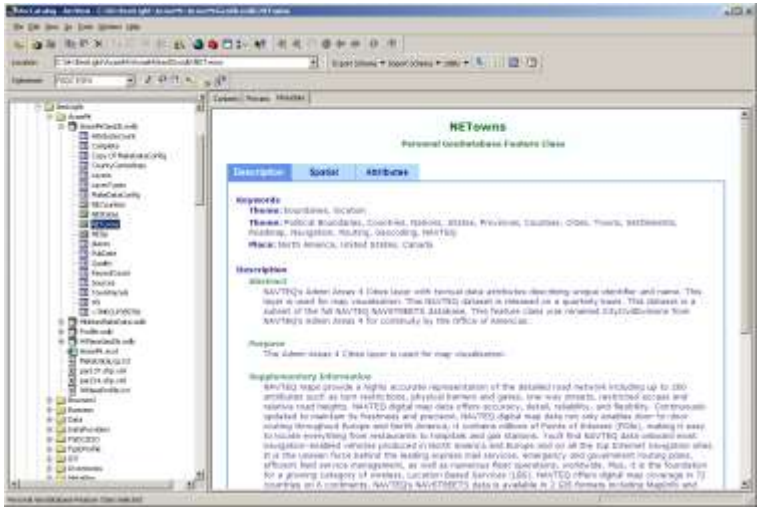
- Finds examples
- Uses ESRI ArcCatalog editor
- Packages with data





Consumers: Using Metadata

- End users: reviews with ArcCatalog

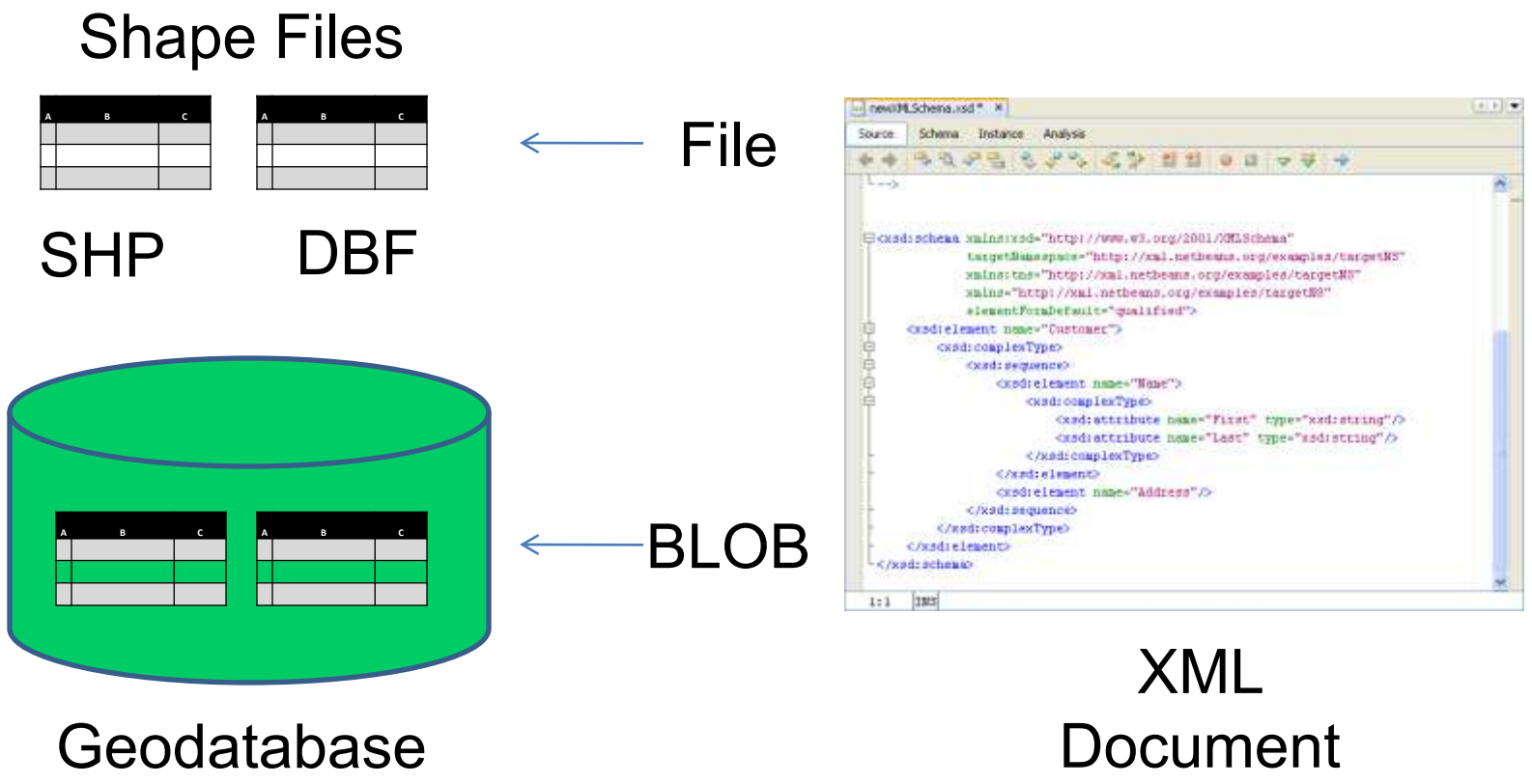


- Applications: Used behind the scenes with GOS, National Map, FGDC, NSDI Nodes

Challenges/Issues

- Learning curve and significant complexity
- Designed to describe single layer, not easy to see a list of all data holdings
- Requires significant manual effort
- Different physical format from data
- Separated from data being described, synchronization problems
- Difficulty representing multi-source data layers
- Difficulty representing feature level metadata
- Entity, Attribute, Domain info is optional
- Domains don’t exist on their own, no name, definition
- Data quality information, percent complete, formats

Geospatial Data and Metadata Separation



Metadata provides an essential ingredient for successful data integration.

Geospatial data and metadata are today managed using very different paradigms. Data is managed using RDBMS and metadata is managed using XML. As such, data and metadata are separated, which leads to synchronization problems and a number of other issues.

Standards Intentions and Physical Implementation

- The FGDC/CSDGM standard clearly states that the physical implementation of metadata management is up to the organization. From the CSDGM Workbook:

The standard specifies information content, ***but not how to organize this information in a computer system or in a data transfer****, or how to transmit, communicate, or present the information to a user. There are several reasons for this approach:*

There are many means by which metadata could be organized in a computer. ***These include incorporating data as part of a geographic information system, in a separate data base, and as a text file.*** *Organizations can choose the approach which suits their data management strategy, budget, and other institutional and technical factors.*

In spite of these clearly stated intentions, metadata implementation has not been organized using GIS or database technology, but rather as separate text/XML/HTML files.

Consider FGDC/ISO XML Metadata output as a database report from the metadata repository, not necessarily as management tool/environment.

Take a Step Back: Why do we develop metadata?

- *Rudyard Kipling:* The "Five Ws" (and one H)
- "Just So Stories" (1902), a poem accompanying the tale of "The Elephant's Child"
- *I keep six honest serving-men*
  (They taught me all I knew);
  *Their names are **What** and **Why** and **When***
  *And **How** and **Where** and **Who***

- Basics: Areas, Categories, Layers, Attributes, Domains

Goal: Maximize Understanding of Data

FGDC Metadata

Data Profiles

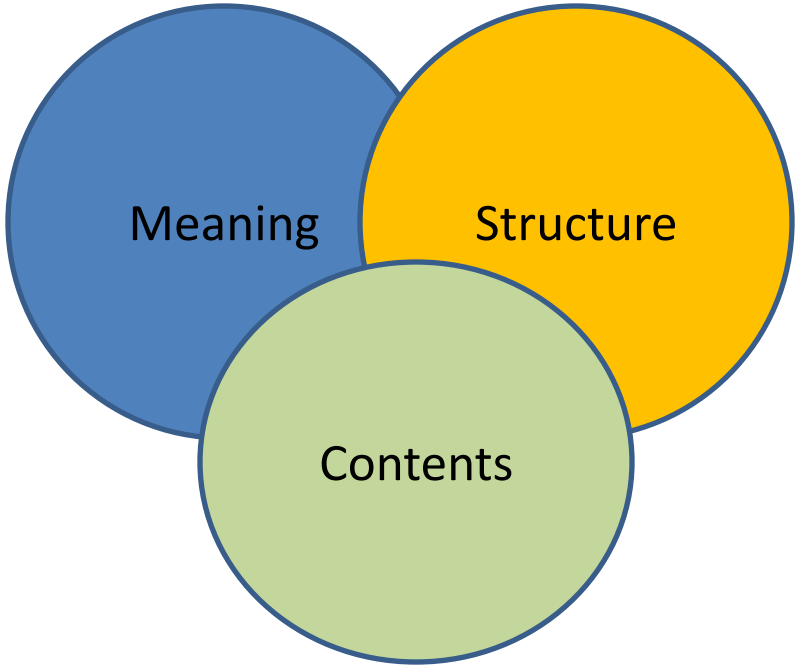
Data Quality Assessments

Cross Referenced Terms

Keywords, Aliases, Indexes

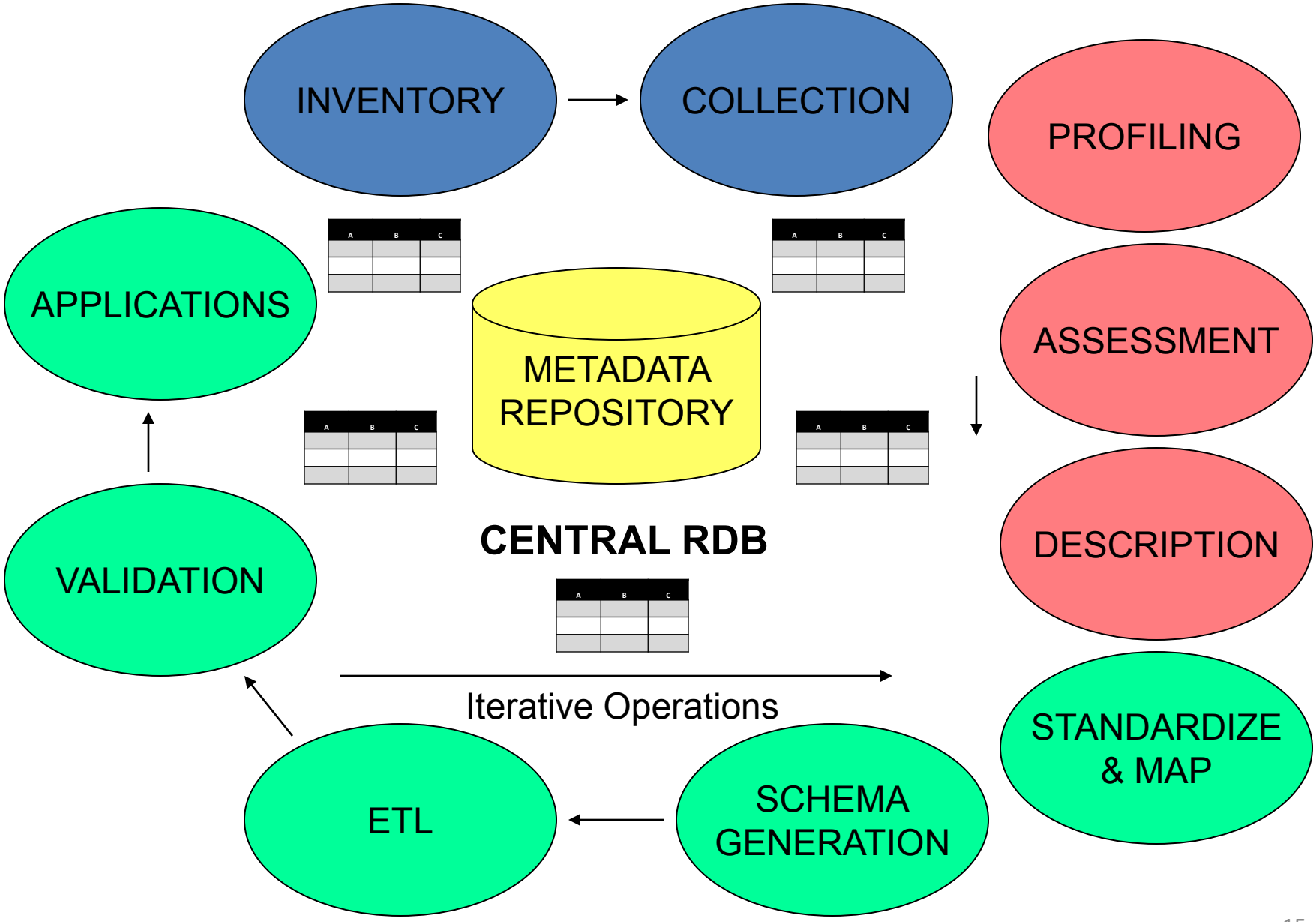
Table of Contents

Glossary



Complete metadata describes Meaning, Structure, and Contents.
Maximize understanding by end user and help write applications.
Help with variety of data description and integration tasks.

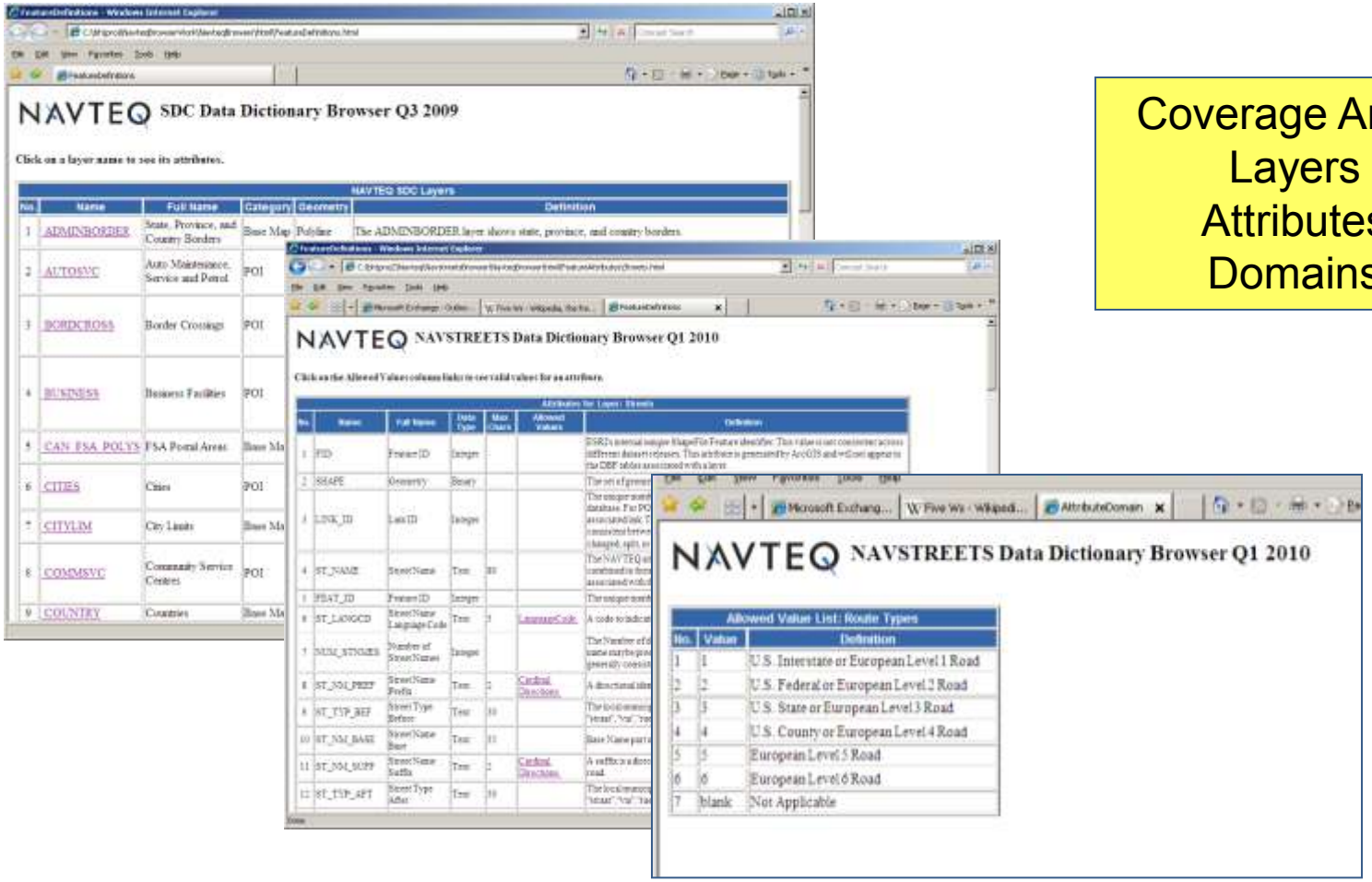
Solution: RDB Metadata Repository and Tools



Benefits: Relational DB Flexibility

- GIS end users intuitively understand data tables
- Useful for describing large, multi-layer, data holdings, easy to see all layers in a list
- RDB applications can be developed for input/output
- Same physical format as data
- Can live alongside data being described
- Can easily represent multi-source data layers
- Can easily represent feature level metadata
- Entity, Attribute, Domain info is cornerstone
- Explicit capture and modeling of domains
- Data quality information, percent complete, formats can be easily integrated
- Facilitates schema matching

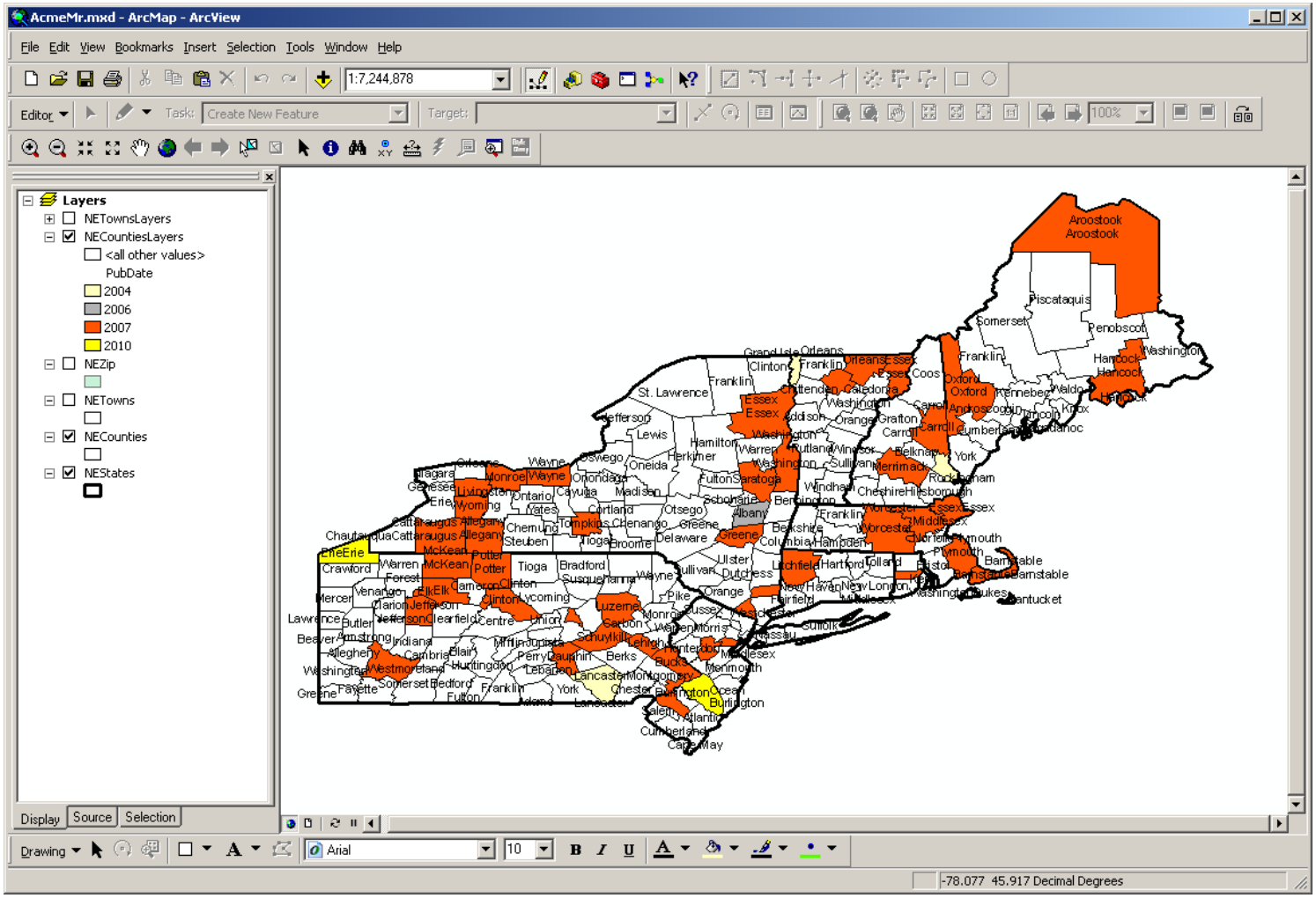
Table Centric Data Dictionaries/Catalogs



Coverage Areas
Layers
Attributes
Domains

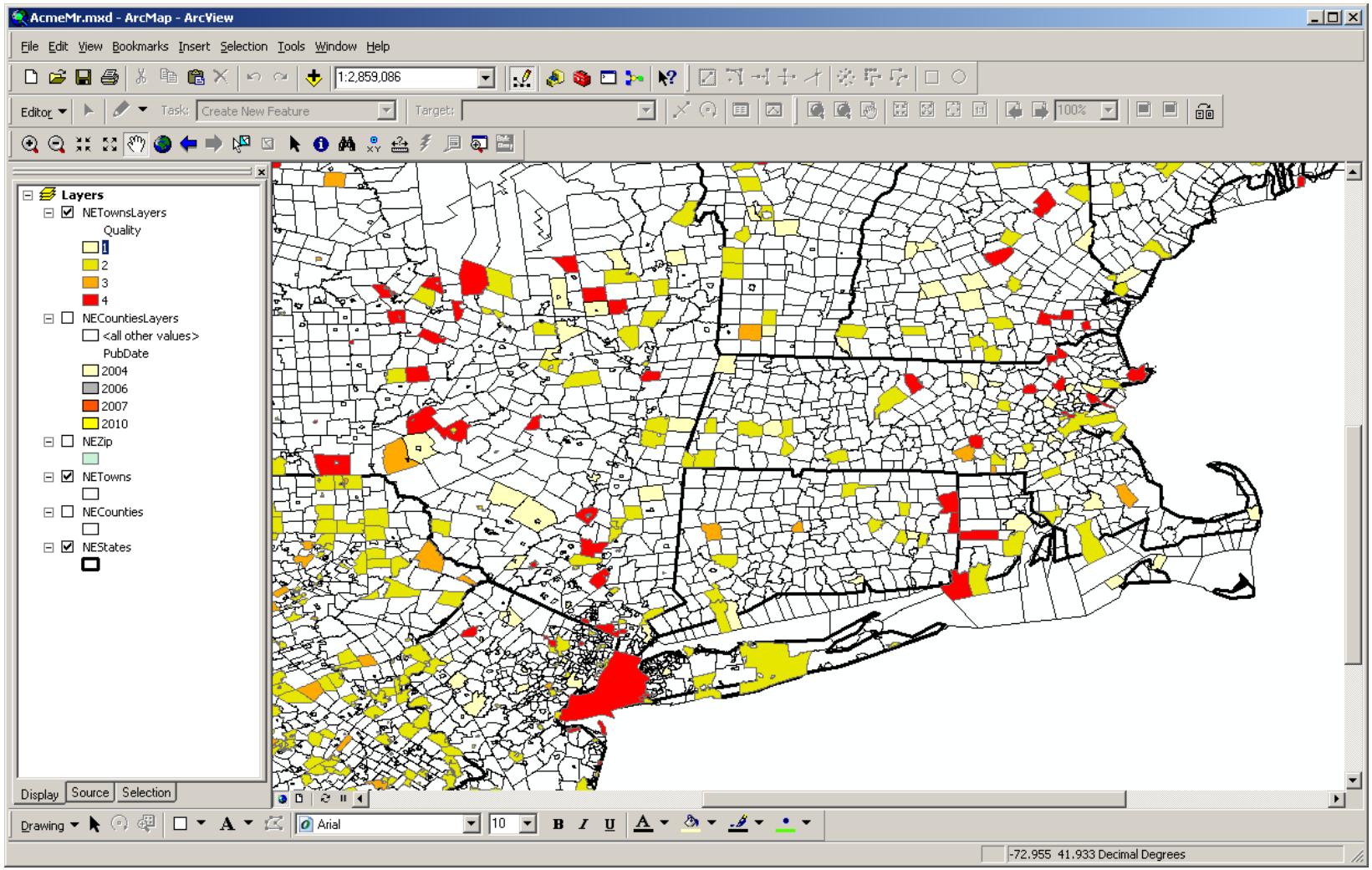
Relational Metadata Repository supports wide variety of useful outputs for end users and application developers.

Meta-Layers: Using Metadata as GIS Data

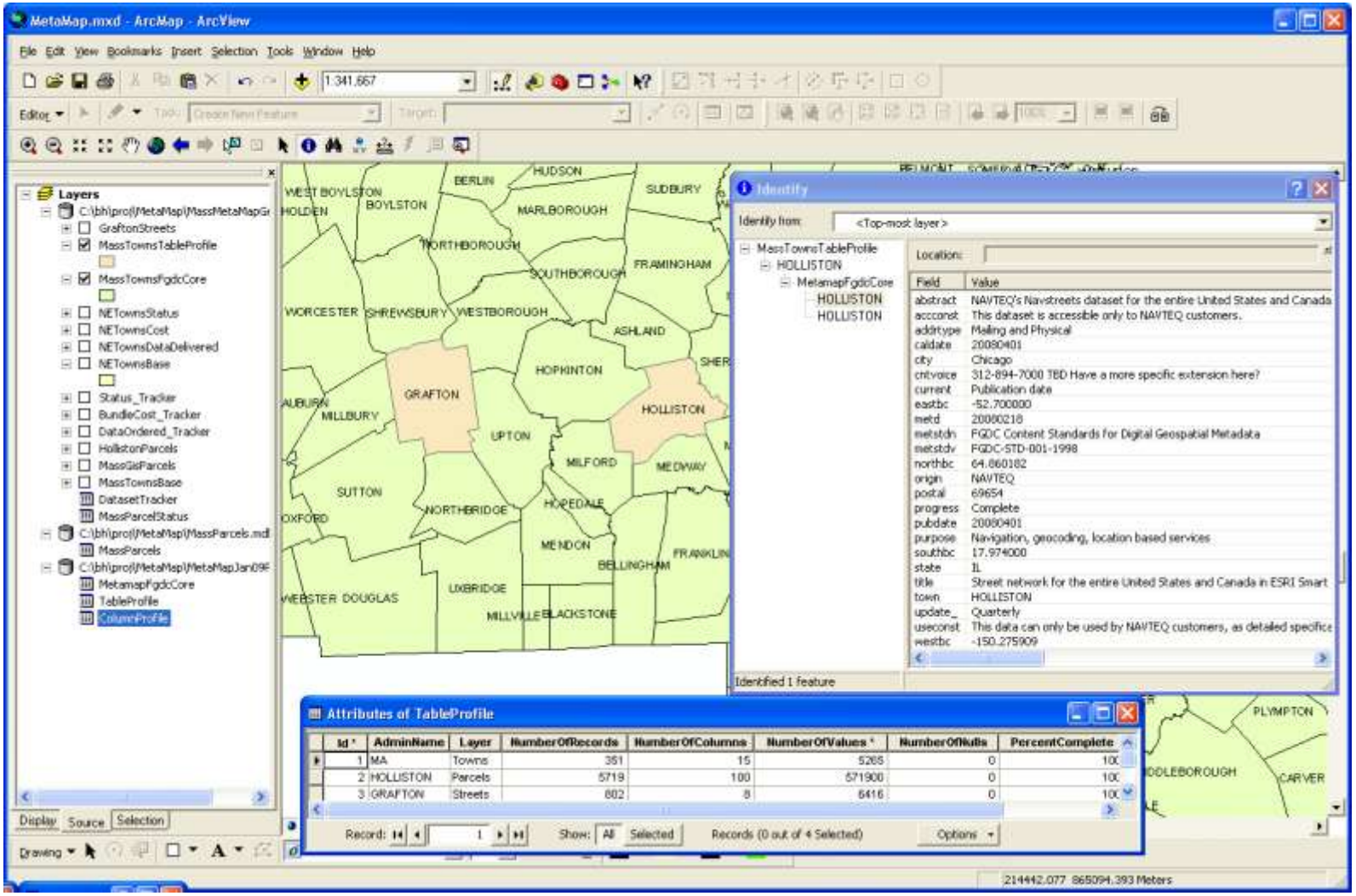


Using metadata the same way we use other GIS data allows wide variety of map presentations, reports, etc. to summarize and highlight datasets by metadata values.

Meta-Layers: Using Metadata as GIS Data (cont.)

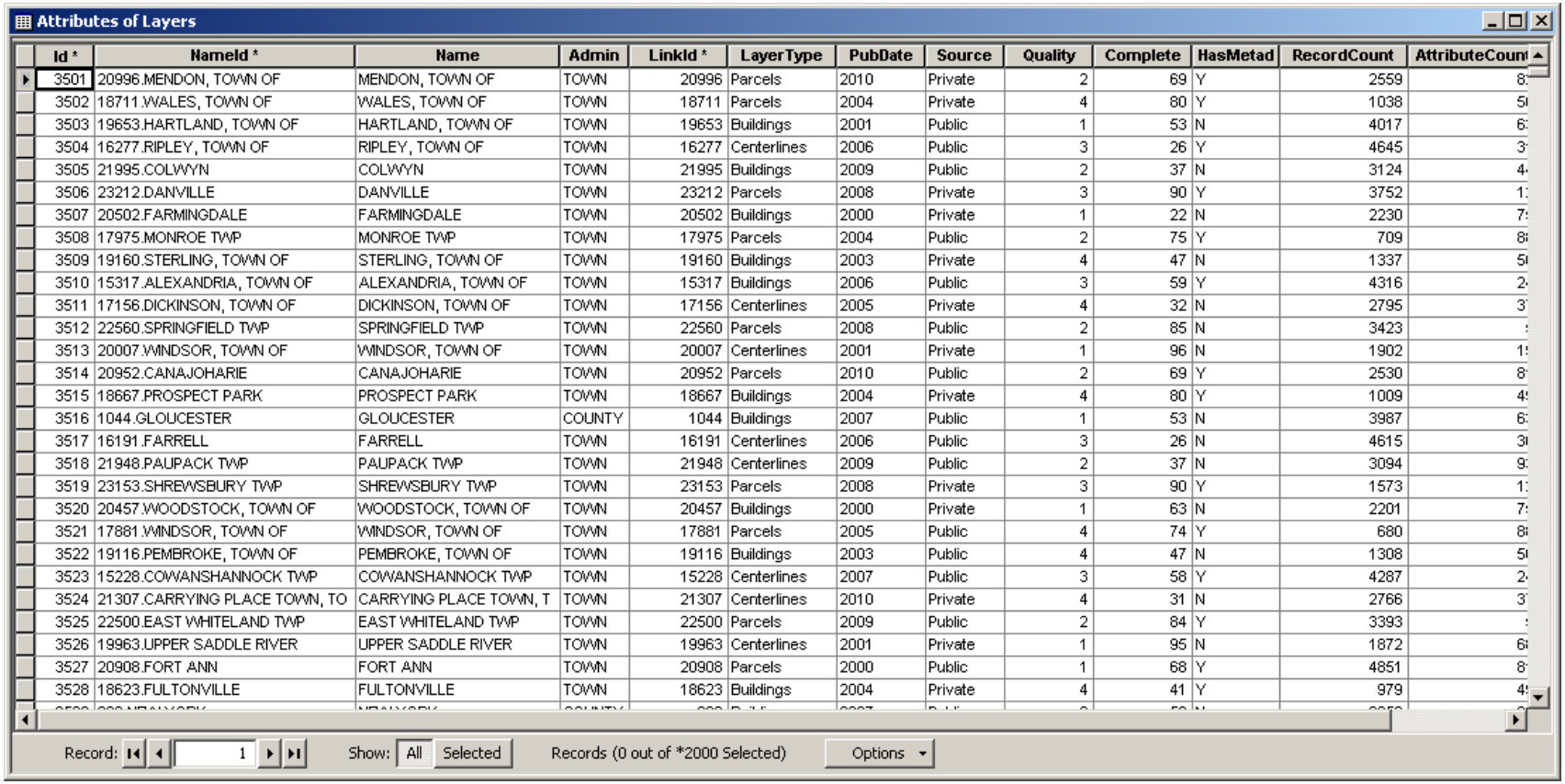


Viewing Meta-Layers in ArcMap



FGDC Metadata is stored in the same RDB, integrated with the geospatial data it is describing, and accessible from ArcMap for review, query, and symbolization.

The Layer-Layer Data



Attributes of Layers

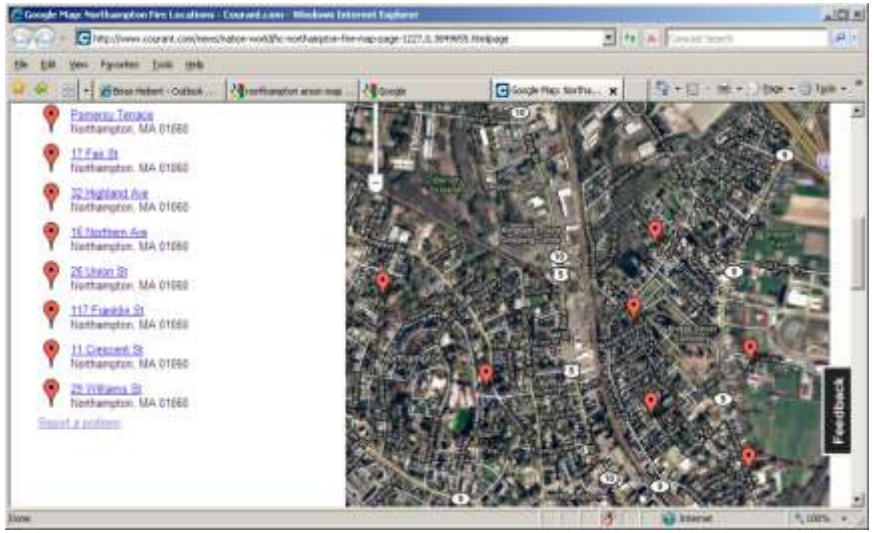
Id *	NameId *	Name	Admin	LinkId *	LayerType	PubDate	Source	Quality	Complete	HasMetad	RecordCount	AttributeCoun
3501	20996.MENDON, TOWN OF	MENDON, TOWN OF	TOWN	20996	Parcels	2010	Private	2	69	Y	2559	8
3502	18711.WALES, TOWN OF	WALES, TOWN OF	TOWN	18711	Parcels	2004	Private	4	80	Y	1038	5
3503	19653.HARTLAND, TOWN OF	HARTLAND, TOWN OF	TOWN	19653	Buildings	2001	Public	1	53	N	4017	6
3504	16277.RIPLEY, TOWN OF	RIPLEY, TOWN OF	TOWN	16277	Centerlines	2006	Public	3	26	Y	4645	3
3505	21995.COLWYN	COLWYN	TOWN	21995	Buildings	2009	Public	2	37	N	3124	4
3506	23212.DANVILLE	DANVILLE	TOWN	23212	Parcels	2008	Private	3	90	Y	3752	1
3507	20502.FARMINGDALE	FARMINGDALE	TOWN	20502	Buildings	2000	Private	1	22	N	2230	7
3508	17975.MONROE TWP	MONROE TWP	TOWN	17975	Parcels	2004	Public	2	75	Y	709	8
3509	19160.STERLING, TOWN OF	STERLING, TOWN OF	TOWN	19160	Buildings	2003	Private	4	47	N	1337	5
3510	15317.ALEXANDRIA, TOWN OF	ALEXANDRIA, TOWN OF	TOWN	15317	Buildings	2006	Public	3	59	Y	4316	2
3511	17156.DICKINSON, TOWN OF	DICKINSON, TOWN OF	TOWN	17156	Centerlines	2005	Private	4	32	N	2795	3
3512	22560.SPRINGFIELD TWP	SPRINGFIELD TWP	TOWN	22560	Parcels	2008	Public	2	85	N	3423	
3513	20007.WINDSOR, TOWN OF	WINDSOR, TOWN OF	TOWN	20007	Centerlines	2001	Private	1	96	N	1902	1
3514	20952.CANAJOHARIE	CANAJOHARIE	TOWN	20952	Parcels	2010	Public	2	69	Y	2530	8
3515	18667.PROSPECT PARK	PROSPECT PARK	TOWN	18667	Buildings	2004	Private	4	80	Y	1009	4
3516	1044.GLOUCESTER	GLOUCESTER	COUNTY	1044	Buildings	2007	Public	1	53	N	3987	6
3517	16191.FARRELL	FARRELL	TOWN	16191	Centerlines	2006	Public	3	26	N	4615	3
3518	21948.PAUPACK TWP	PAUPACK TWP	TOWN	21948	Centerlines	2009	Public	2	37	N	3094	9
3519	23153.SHREWSBURY TWP	SHREWSBURY TWP	TOWN	23153	Parcels	2008	Private	3	90	Y	1573	1
3520	20457.WOODSTOCK, TOWN OF	WOODSTOCK, TOWN OF	TOWN	20457	Buildings	2000	Private	1	63	N	2201	7
3521	17881.WINDSOR, TOWN OF	WINDSOR, TOWN OF	TOWN	17881	Parcels	2005	Public	4	74	Y	680	8
3522	19116.PEMBROKE, TOWN OF	PEMBROKE, TOWN OF	TOWN	19116	Buildings	2003	Public	4	47	N	1308	5
3523	15228.COWANSHANNOCK TWP	COWANSHANNOCK TWP	TOWN	15228	Centerlines	2007	Public	3	58	Y	4287	2
3524	21307.CARRYING PLACE TOWN, TO	CARRYING PLACE TOWN, T	TOWN	21307	Centerlines	2010	Private	4	31	N	2766	3
3525	22500.EAST WHITELAND TWP	EAST WHITELAND TWP	TOWN	22500	Parcels	2009	Public	2	84	Y	3393	
3526	19963.UPPER SADDLE RIVER	UPPER SADDLE RIVER	TOWN	19963	Centerlines	2001	Private	1	95	N	1872	6
3527	20908.FORT ANN	FORT ANN	TOWN	20908	Parcels	2000	Public	1	68	Y	4851	8
3528	18623.FULTONVILLE	FULTONVILLE	TOWN	18623	Buildings	2004	Private	4	41	Y	979	4

Record: 1 Show: All Selected Records (0 out of *2000 Selected) Options

Meta-Layer data table includes information describing dataset: coverage area, level of political administration, layer contents, source, quality, publication date, completeness, whether or not FGDC metadata is available, number of records, number of attributes, etc.

Record Level Metadata

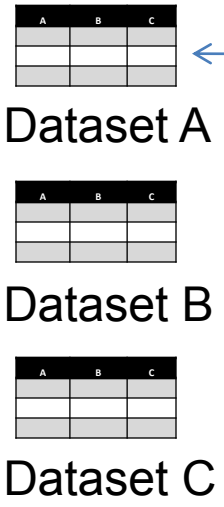
- Geospatial metadata describes the group of records comprising a dataset as a single entity.
- Some end users want metadata at the individual record level.
- FGDC/ISO does not support this granularity.



Name	Value
Contact How:	Telephone
Contact Date	11-May-10
Location Confirmed	Y
Moved Geocoded	Y
Accuracy	Building Footprint
Notification	N
Site Image	N
Status	Closed

Multi-Source Data Layers

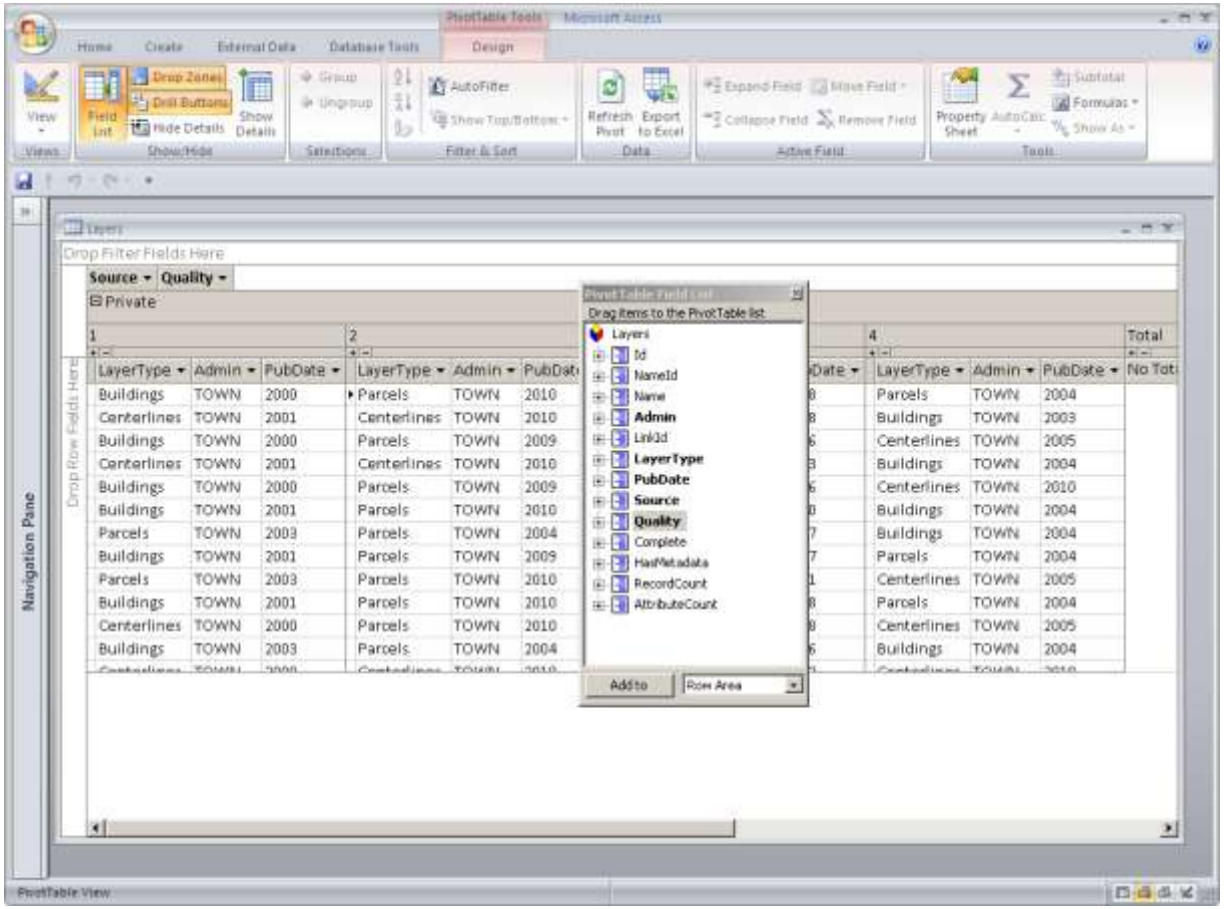
- Some data layers are the result of a merge between multiple input layers.
- To track this, each record needs to carry a link back to the original dataset.
- Necessary if swap-out updates are part of workflow.
- New ISO 19115 has notion of dataset series.



Name	Value
ObjectId	88197
SOURCE ID	**134**
Name	Acme GIS
Address1	100 Elm St.
Address2	Suite 100
City	Northampton
State	MA
Zip	10160

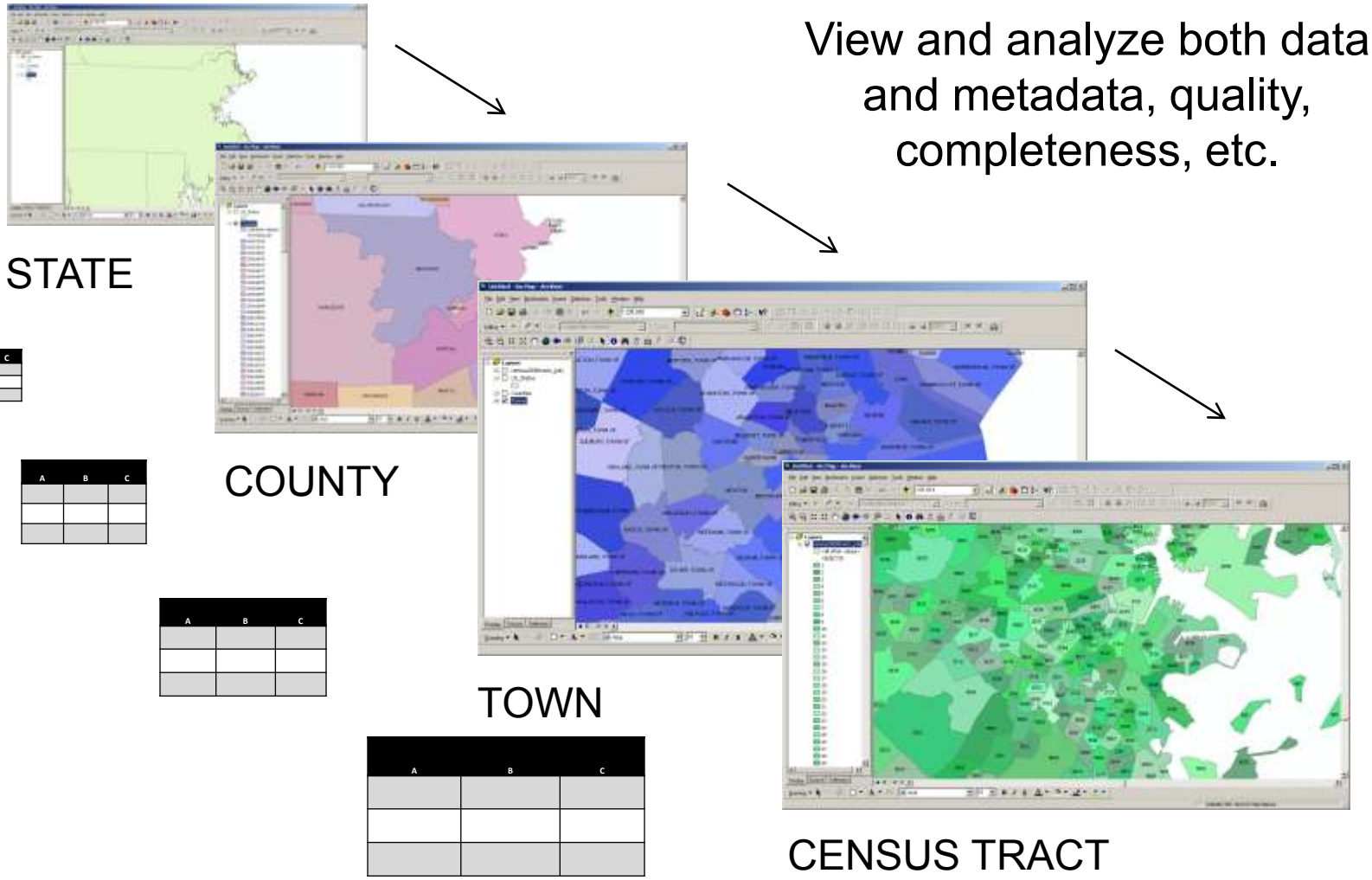
Pivot Tables: Standard Biz-Intel Tool

- Hierarchical
- Drill-down
- Aggregates
- Easy to Use



Sample Pivot Table in MS_Access showing Layer Type, Admin Area, Publication Date, Source, Quality

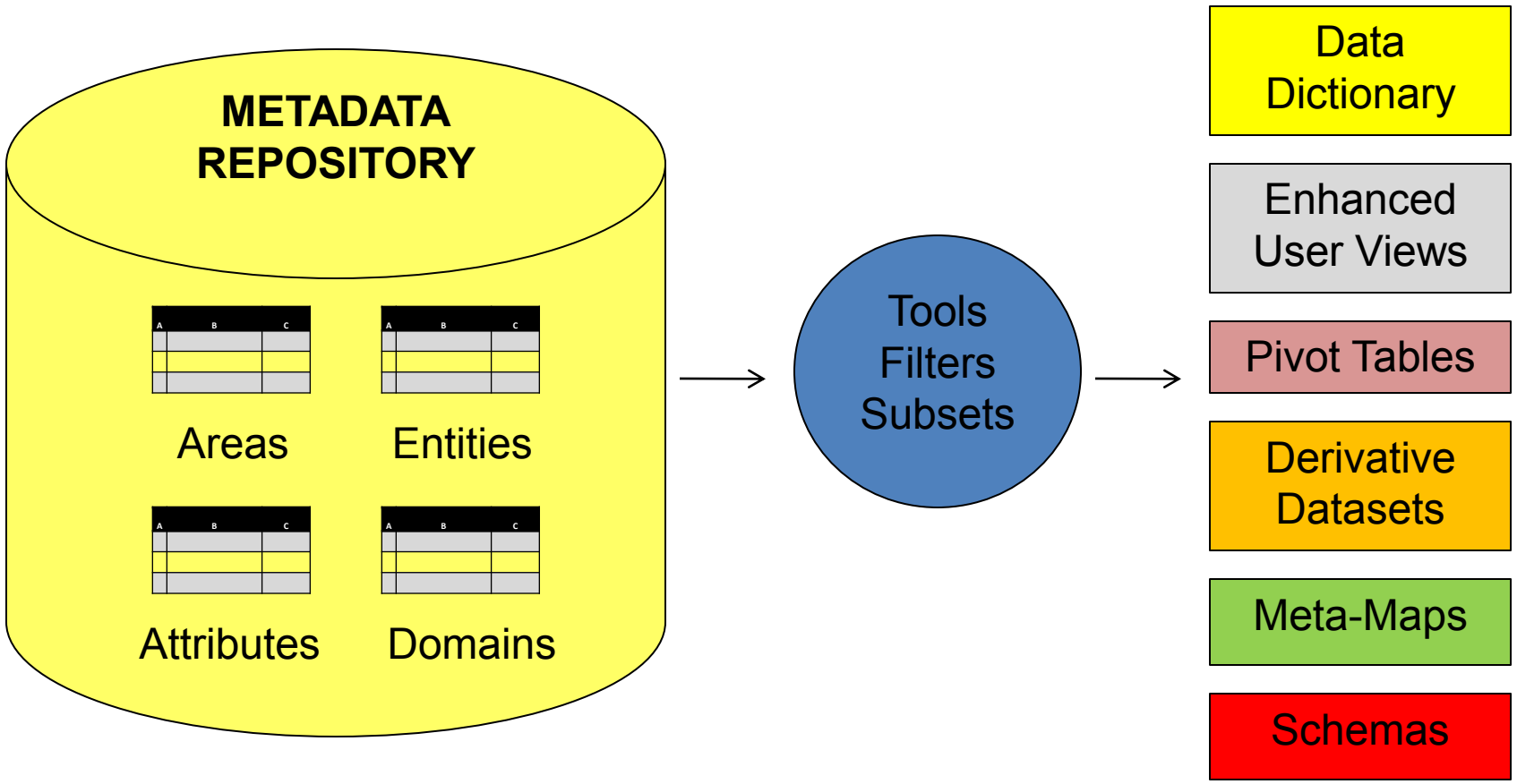
Future: Business/Geo-Intelligence Pivot Tables/Maps



View and analyze both data and metadata, quality, completeness, etc.

Business Intelligence data exploration/viewing solutions make heavy use of **pivot tables, drill-down, drill-through**. With a data warehousing approach, geospatial intelligence solutions can use a similar approach, with maps

Metadata Repository Outputs

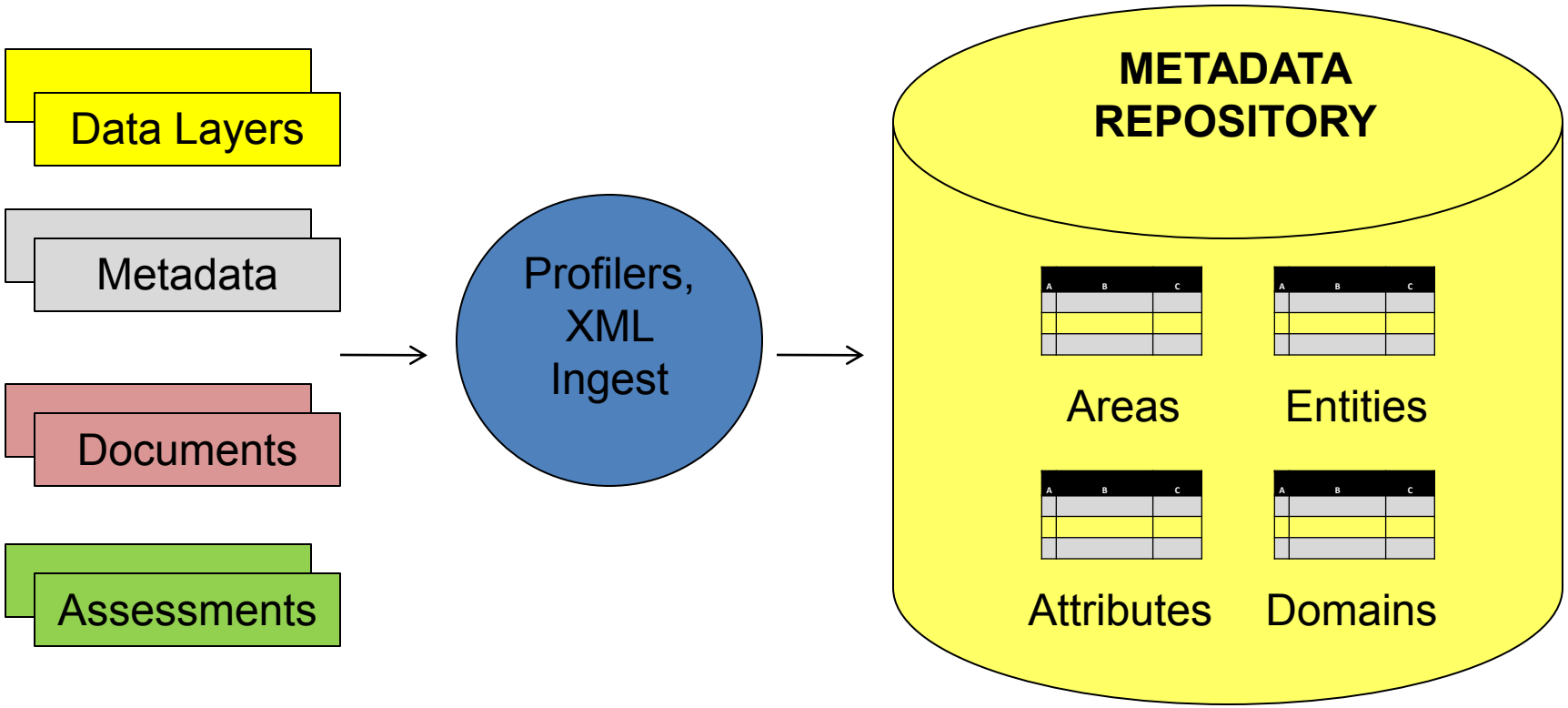


Metadata Repository and tools can support the generation of a wide variety of custom data and metadata products

How to Build the Geospatial Metadata Repository: A RDB Data Model and Input/Output Tools

- **Data Model**: Tables for Areas, Categories, Layers, Attributes, Domains
- Collect existing data, metadata, documentation, knowledge
- Use **data profiling** tool to capture basic schema structure/contents
- Use **XML metadata ingest** tool to help capture meaning
- Use **Meta-Layer Geometry** tools to create or associate metadata with geometric features representing dataset areas

Loading the Geospatial Metadata Repository



The Metadata Repository, implemented as an RDMBS, is populated through a combination of tools and manual reviews/assessments.

The Data Profile

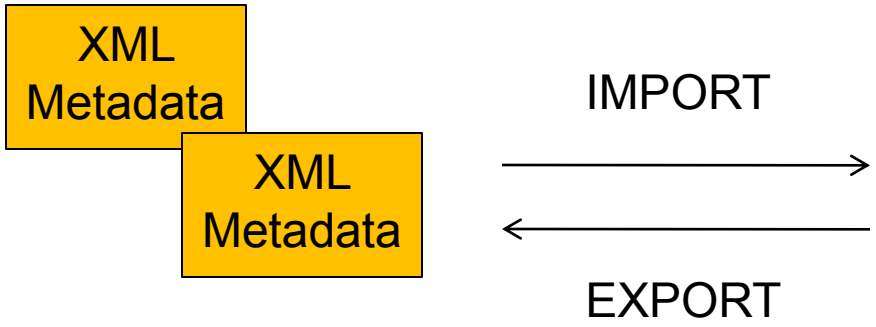
NUM	ELEMENT	DEFINITION
1	DatasetId	A uniqe identifier for the dataset
2	DatabaseName	The name of the source database
3	TableName	The name of the source database table
4	RecordCount	The number of records in the table
5	ColumnCount	The number of columns in the table
6	NumberOfNulls	The number of null values in the table

The Table Profile is helpful for getting a good overall idea of what's in a database

The Column Profile is helpful for getting a detailed understanding of database structure and contents

NUM	ELEMENT	DEFINITION
1	DatasetId	A uniqe identifier for the dataset
2	DatabaseName	The name of the database
3	TableName	The name of the database table
4	ColumnName	The name of the data column
5	DataType	The data type of the column
6	MaxLength	The max length of the column
7	DistinctValues	The number of distinct values used in the column
8	PercentDistinct	The percentage of distinct values used in the column
9	SampleValues	A sampling of data values used in the column
10	MinLengthValue	The minimum length data value
11	MaxLengthValue	The maximum length data value
12	MinValue	The minimum value
13	MaxValue	The maxim value

FGDC/ISO XML Metadata and the RDB



When this metadata is imported into an RDB, the full flexibility of SQL is available for very flexible management and querying a large collection of metadata as a set.

It's easy to exchange data between XML and RDB

NUM	ELEMENT
1	Originator
2	Publication_Date
3	Title
4	Abstract
5	Purpose
6	Calendar_Date
7	Currentness_Reference
8	Progress
9	Maintenance_and_Update_Frequency
10	West_Bounding_Coordinate
11	East_Bounding_Coordinate
12	North_Bounding_Coordinate
13	South_Bounding_Coordinate
14	Theme_Keyword_Thesaurus
15	Theme_Keyword
16	Access_Constraints
17	Metadata_Date
18	Contact_Person
19	Address_Type
20	Address
21	City
22	State_or_Province
23	Postal_Code
24	Contact_Voice_Telephone
25	Metadata_Standard_Name
26	Metadata_Standard_Version

XML Metadata After Import into RDB: Hierarchy Preserved

Name	NodeValue	ParentId	ParentName	LineageId	LineageName
origin	ACME	5	citeinfo	1.2.3.4.6	metadata.idinfo.citation.citeinfo.pubdateX
pubdate	05/21/2004	5	citeinfo	1.2.3.4.7	metadata.idinfo.citation.citeinfo.title
title	Centerlines	5	citeinfo	1.2.8	metadata.idinfo.descript
geoform	vector digital data	5	citeinfo	1.2.8.10	metadata.idinfo.descript.purpose
ftname	cntrline	5	citeinfo	1.2.11.12.13.14	metadata.idinfo.timeperd.timeinfo.sngdate.caldateX
abstract	This dataset …	17	descript	1.2.16.18	metadata.idinfo.status.update
purpose	Navigation	17	descript	1.2.19	metadata.idinfo.spdom
langdata	en	17	descript	1.2.19.20	metadata.idinfo.spdom.bounding
caldate	20040528	24	sngdate	1.2.25	metadata.idinfo.keywords
time	unknown	24	sngdate	1.2.25.26	metadata.idinfo.keywords.theme
current	publication date	22	timeperd	1.2.25.26.27	metadata.idinfo.keywords.theme.themekt
progress	In work	28	status	1.2.29	metadata.idinfo.accconst
update	Unknown	28	status	1.2.30	metadata.idinfo.useconst
westbc	-178.047715	32	bounding	1.31.33	metadata.metainfo.metc
eastbc	174.060288	32	bounding	1.31.33.34.35	metadata.metainfo.metc.cntinfo.cntperp
northbc	65.634111	32	bounding	1.31.33.34.37	metadata.metainfo.metc.cntinfo.cntaddr
southbc	17.650000	32	bounding	1.31.33.34.37.39	metadata.metainfo.metc.cntinfo.cntaddr.address
themekey	Geocoding, Routing	51	theme	1.3.50.51.52	metadata.idinfo.keywords.theme.themekey

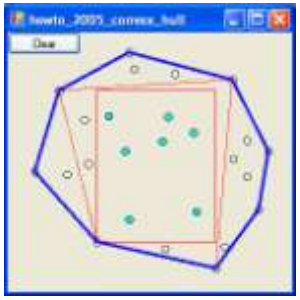
Meta-Layer Geometry Creation and Management

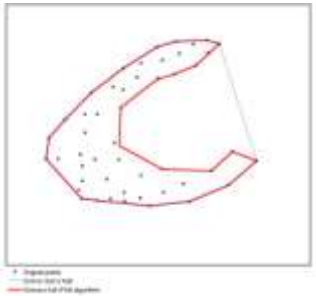
1

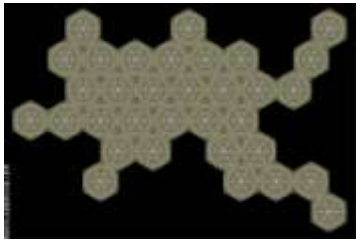
Lon/Lat Bounding Boxes

```
Spatial_Domain:
  Bounding_Coordinates:
    West_Bounding_Coordinate: -167.946360
    East_Bounding_Coordinate: 179.001991
    North_Bounding_Coordinate: 71.298141
    South_Bounding_Coordinate: 17.678360
```

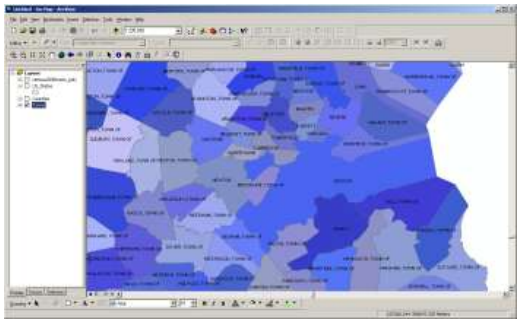
2





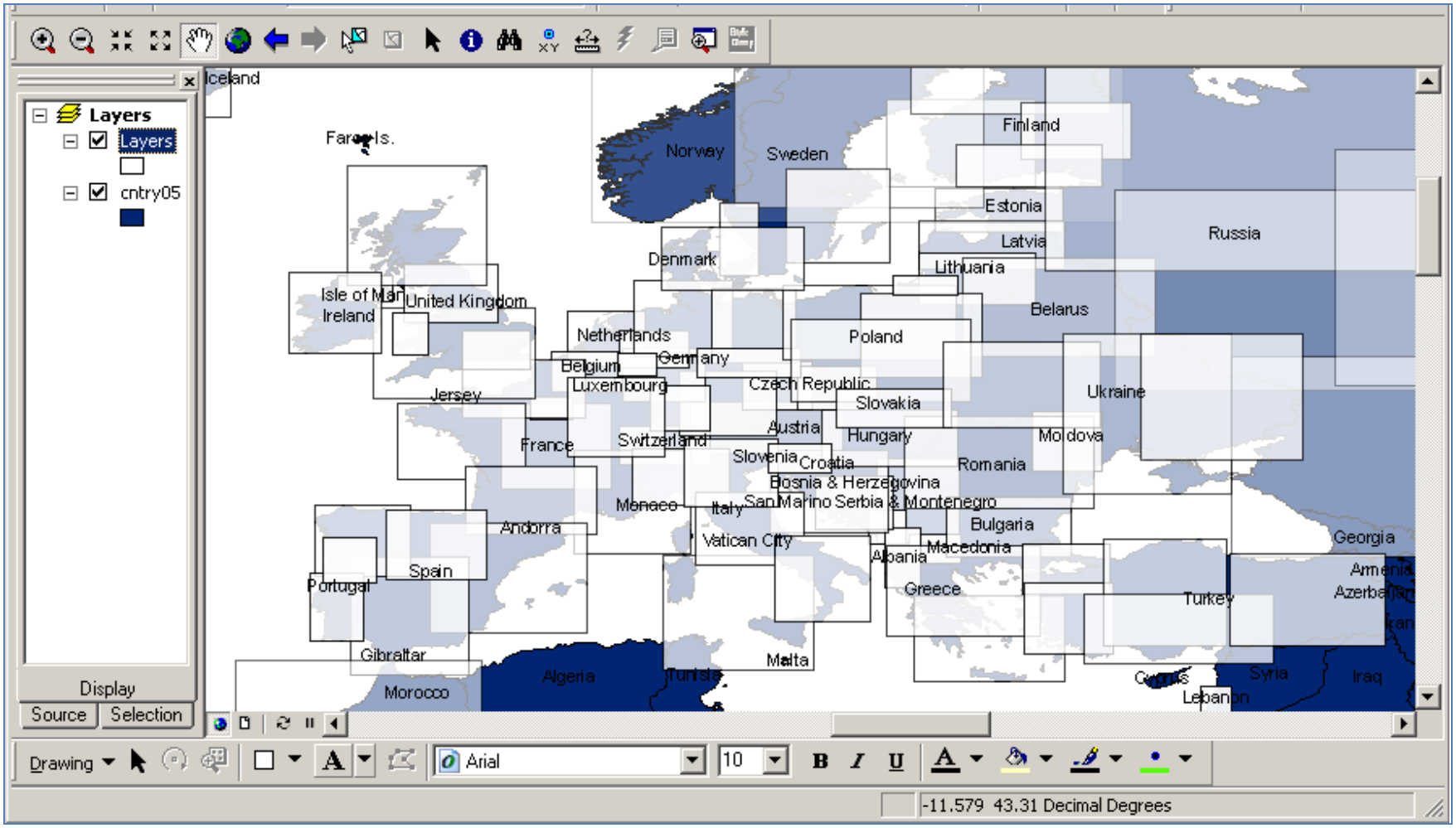


3



Three basic approaches to generating layer coverage polygons with increasing level-of-effort as 1) bounding boxes 2) convex/concave hulls, tessellations and 3) existing administrative or other polygons. Choice based on presentation and data management requirements.

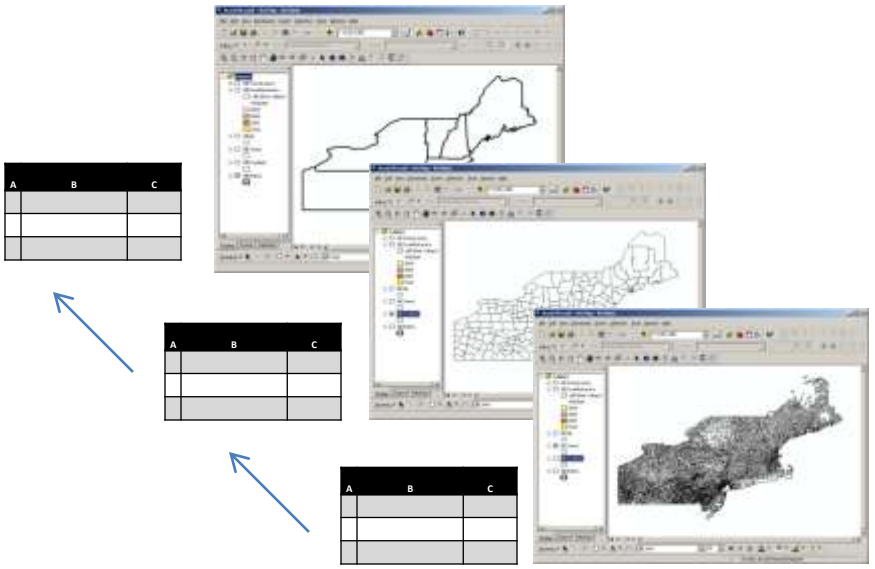
Meta-Layer Dataset Outlines using Boxes

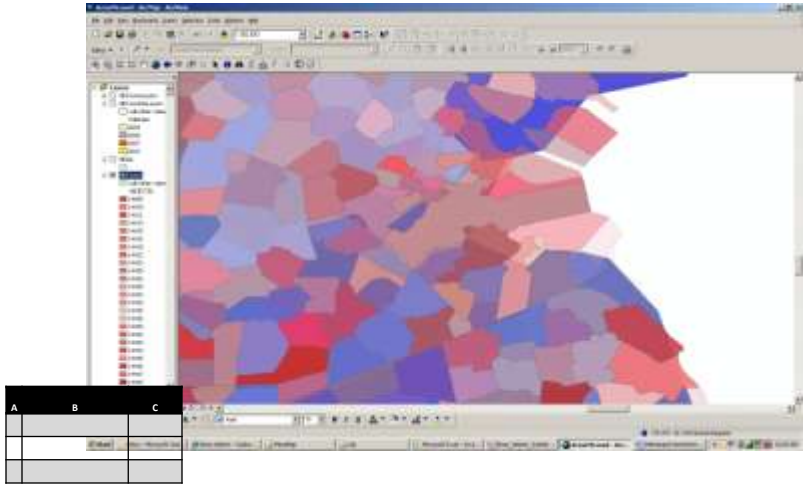


Geospatial data provider coverage of Europe using bounding boxes for meta-layer dataset outlines.

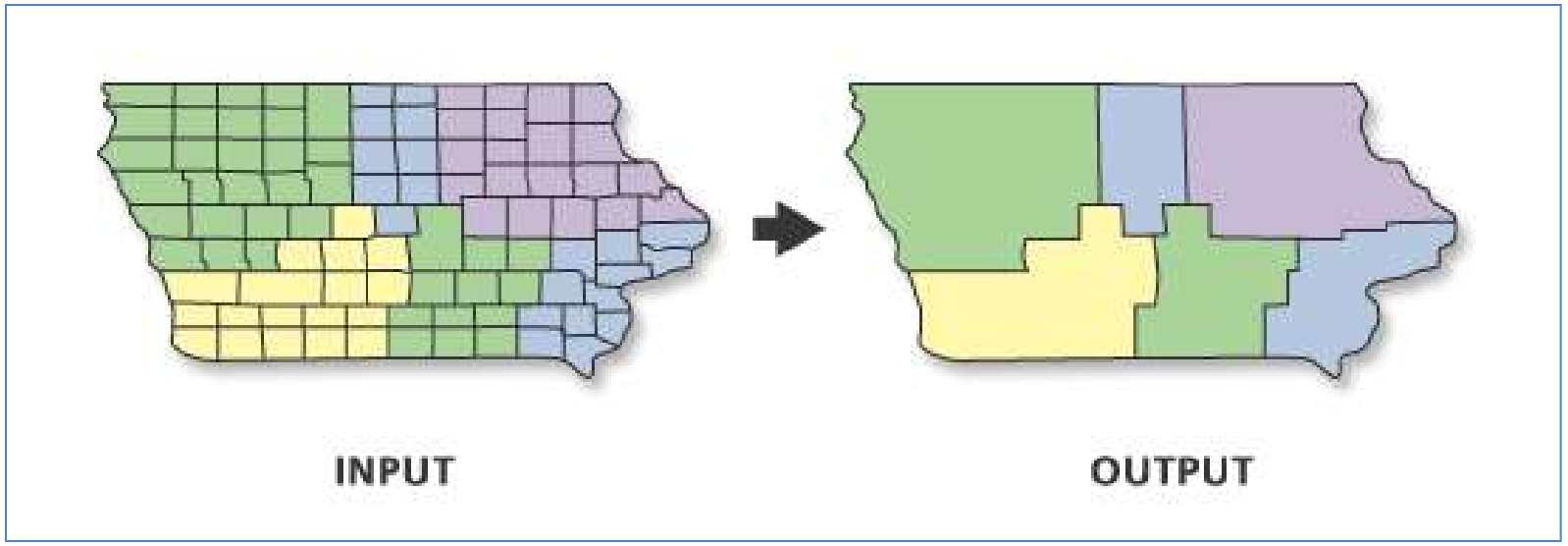
Meta-Layer Dataset Outlines using Existing Geography

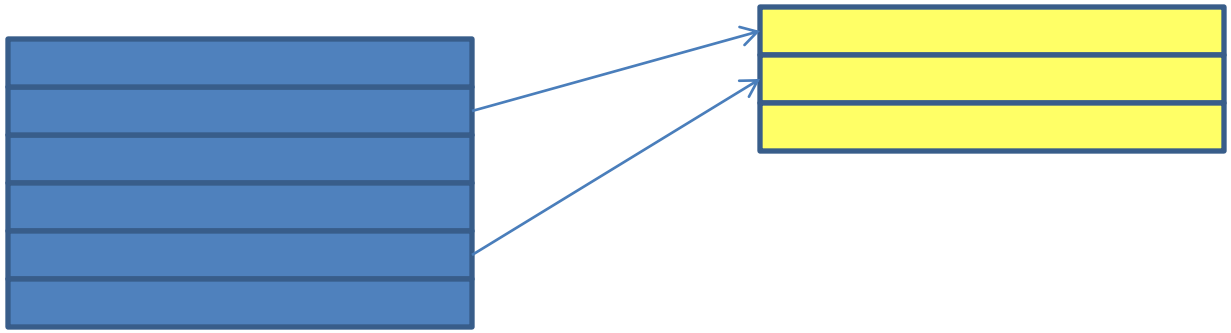
- Need to track/manage multi-level hierarchy?
- Need to capture notion of partial coverage using subset elements, e.g., how many towns per county available?
- How many meta-layers?
- Multiple tables or single parent/child, recursive table?
- Presentation, transparency, reporting?
- Level of generalization?





Use of ArcMap Merge/Dissolve to Create Aggregate Dataset Polygons



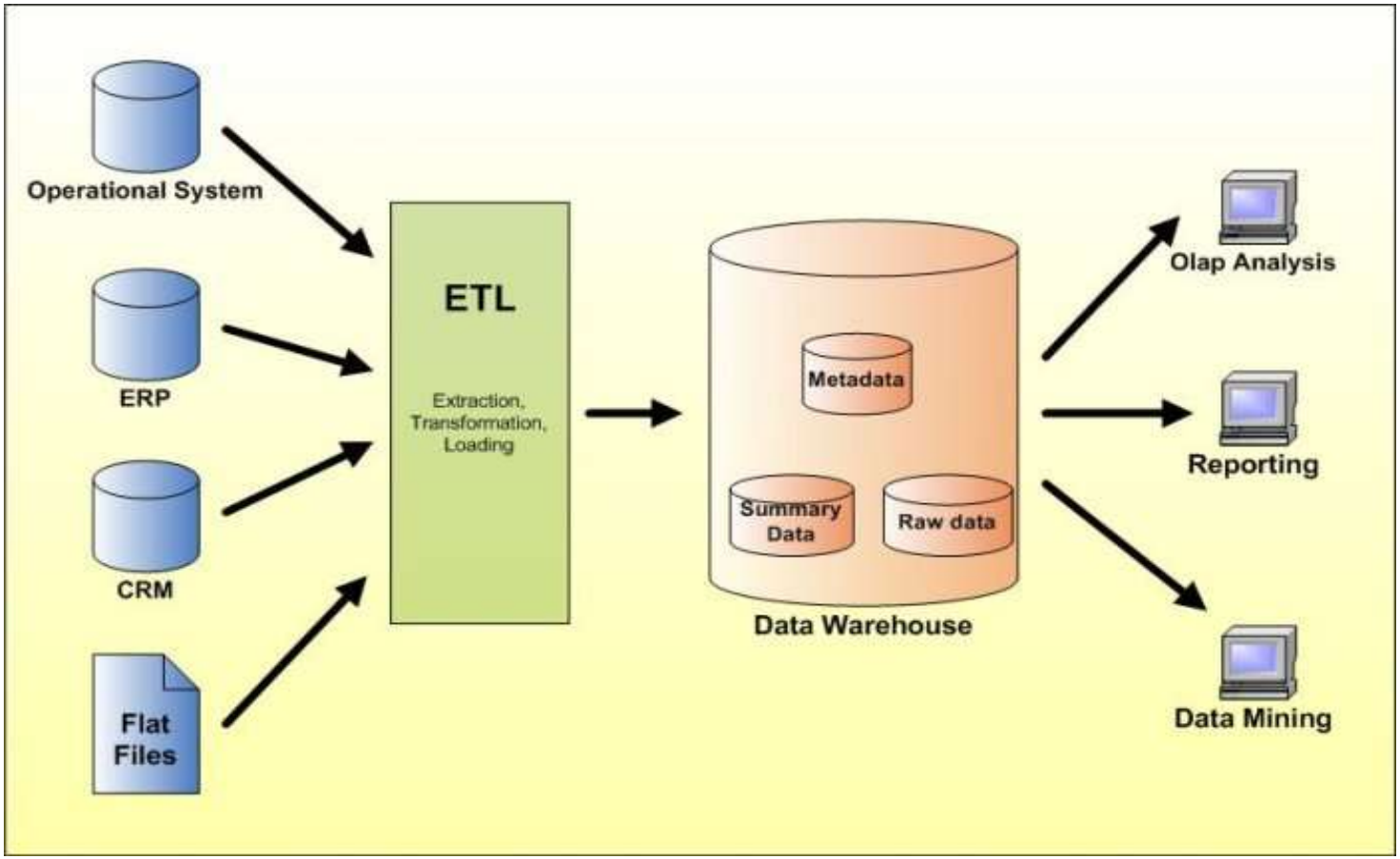


Using manual or automated techniques with custom ArcObjects applications, partial dataset coverage, in a hierarchical context, can be managed/represented using Merge/Dissolve routines and related records.

IT and Geospatial Paradigms and Cultures: Tools and Techniques to Solve Problems

- Decision Support, **Data Warehousing**
- Structured vs. Unstructured Data Access
- **Appropriate use of standards for mission goals**
- Information Retrieval, Library Science, Semantic Web
- Spatial **Ontology** Community of Practice (SOCoP)
- Business Intelligence, ETL, OLAP Cubes
- AGILE Software Development
- **Data Quality, Profiling**
- Metadata Repositories, QA/QC
- Business Process Management
- Lean Manufacturing

Data Warehousing and Metadata Repositories



To support business intelligence and decision support systems, mainstream IT data integration makes heavy use of relational and multi-dimensional data warehouses and **metadata repositories**.

Data Quality Book

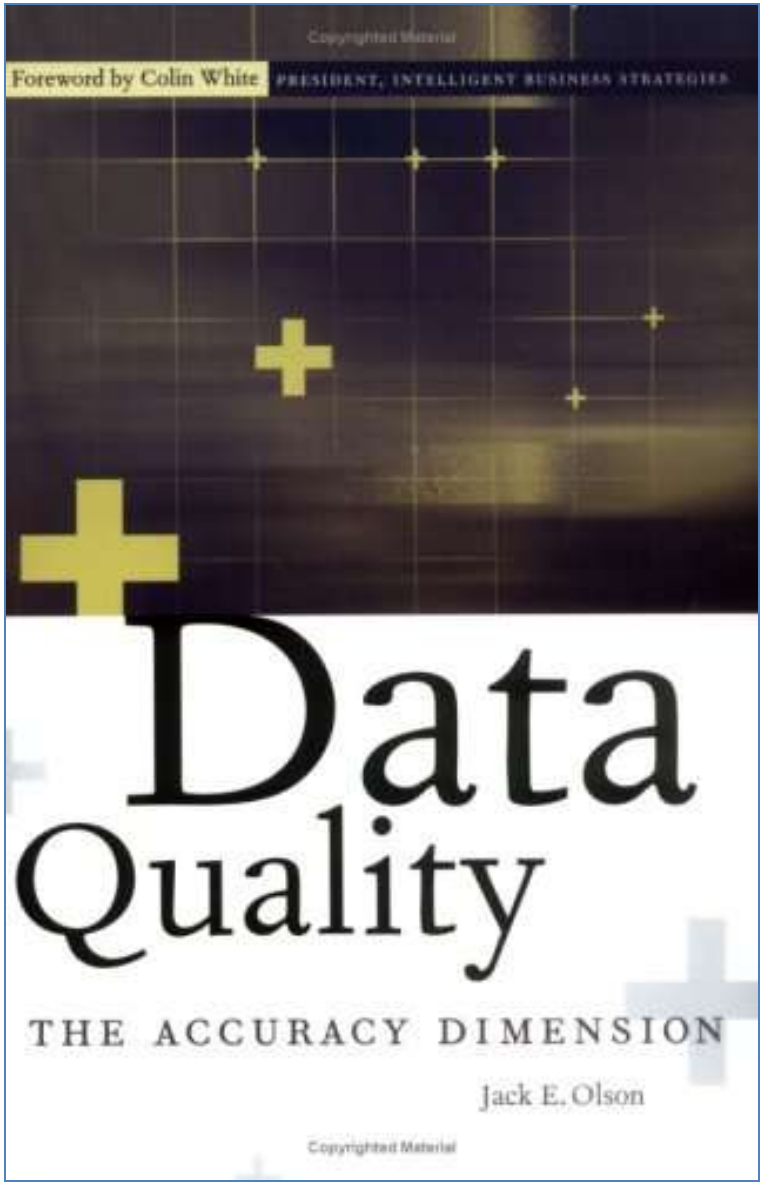
Data Quality: The Accuracy Dimension

Jack E. Olson

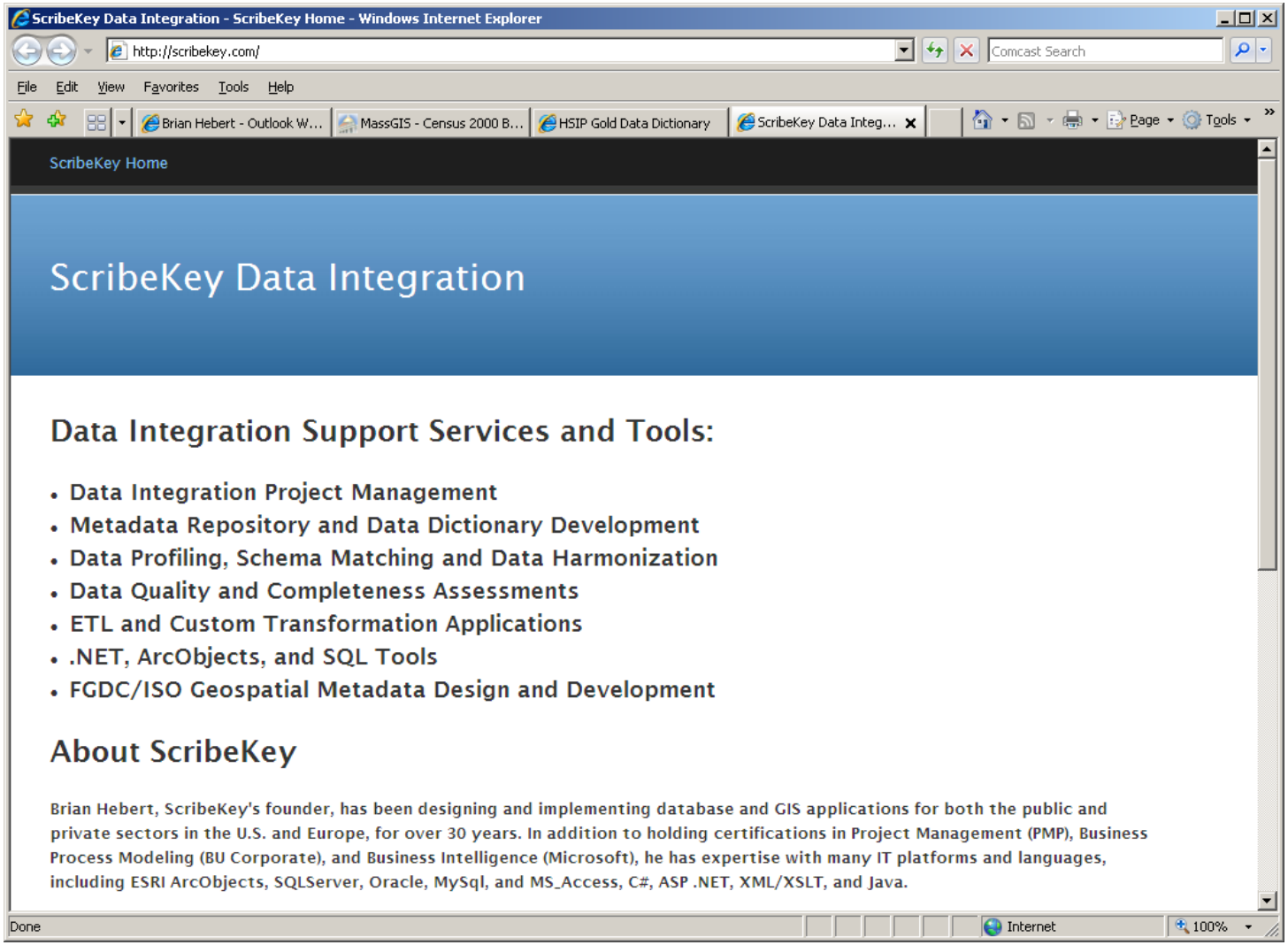
Morgan Kaufmann; 1 edition (January 9, 2003)

ISBN-10: 1558608915
ISBN-13: 978-1558608917

Discusses profiling and metadata repository in great detail.



www.scribekey.com



Take Aways:

Geospatial Metadata **IS** Geospatial Data

Leverage Desktop/Web DB/GIS Technology

Use Standards Appropriately

Thank You

Questions and Answers

bhebert@scribekey.com